11/05/2020

MVA ASSIGNMENT 8

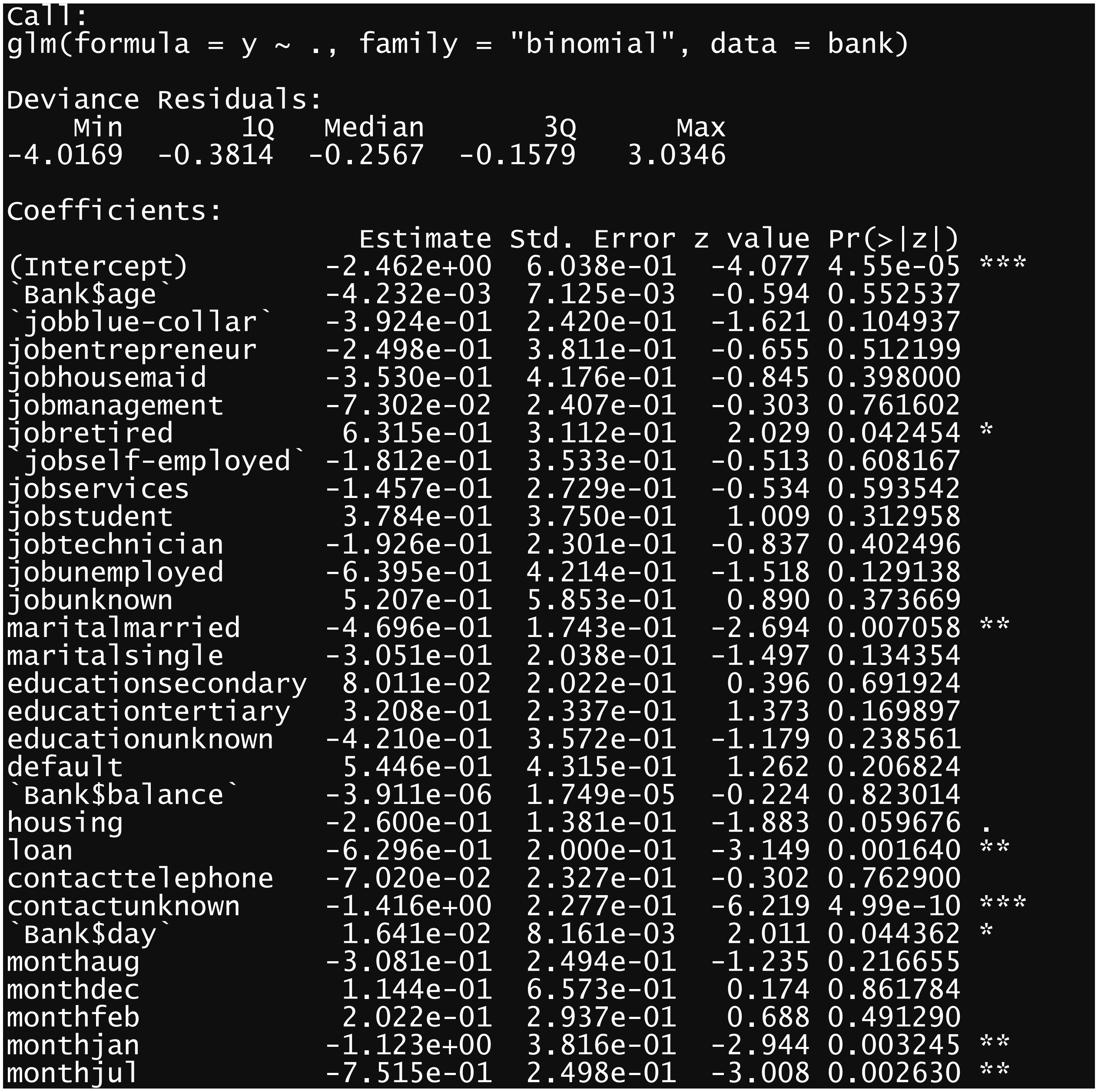
MEMBER INFORMATION

Ruixin Yang (RUID: 197000459)

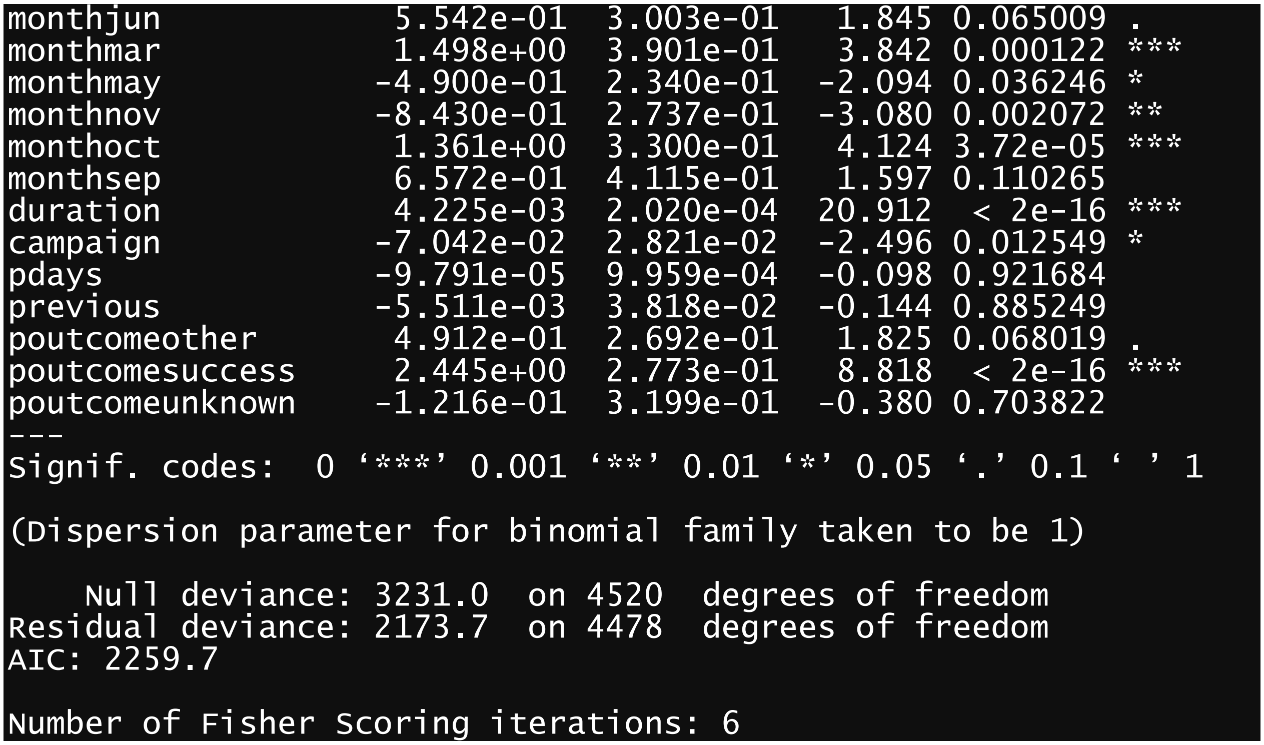
Aishwarya Senthilvel (RUID: 199001269)

LOGISTIC REGRESSION ANALYSIS

First, we ran the overall logistic regression to all the variables with the outcome of purchasing the deposit product as our dependent binary variable. Here is the outcome.

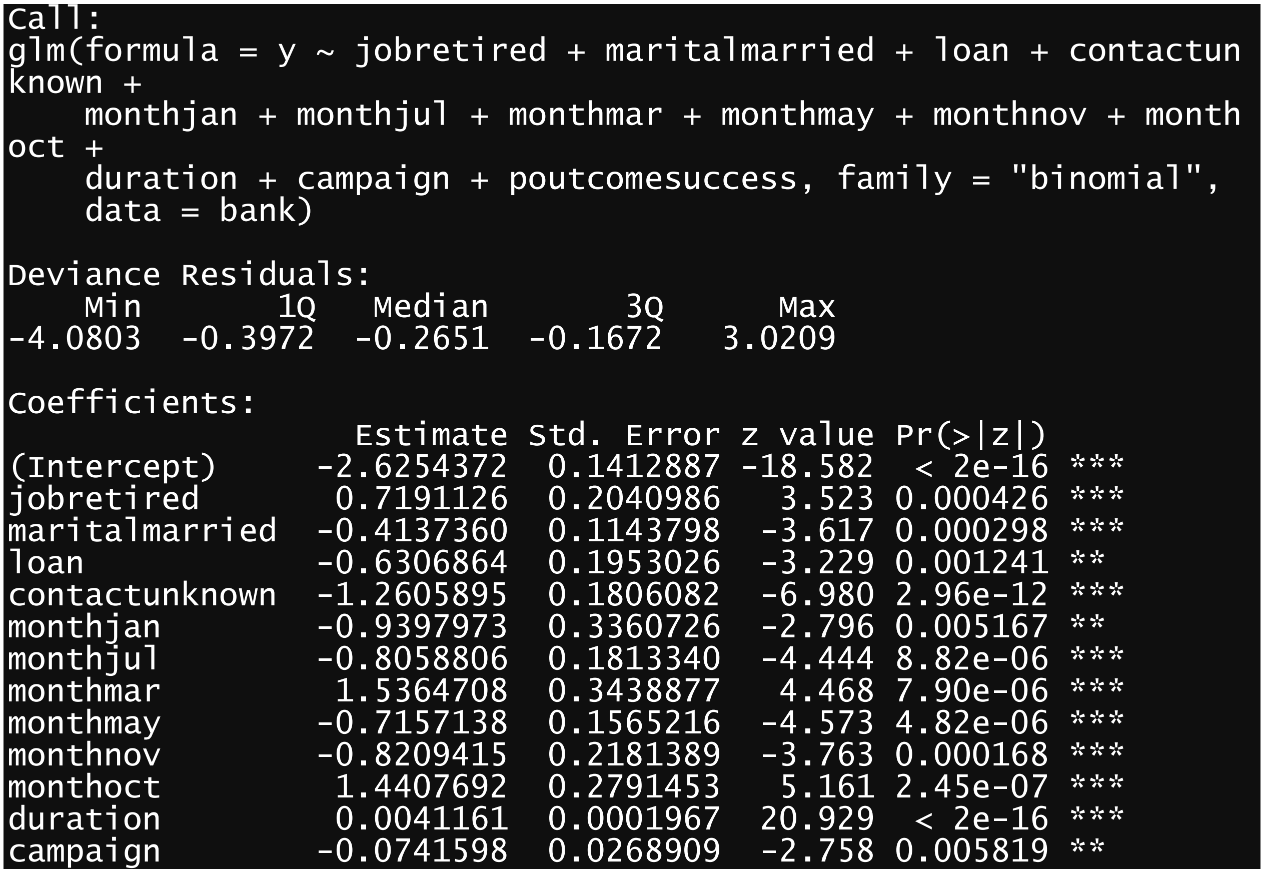


```
Call:
glm(formula = y ~ ., family = "binomial", data = bank)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.0169  -0.3814  -0.2567  -0.1579   3.0346

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)        -2.462e+00  6.038e-01  -4.077 4.55e-05 ***
`Bank$age`         -4.232e-03  7.125e-03  -0.594 0.552537
`jobblue-collar`   -3.924e-01  2.420e-01  -1.621 0.104937
jobentrepreneur    -2.498e-01  3.811e-01  -0.655 0.512199
jobhousemaid       -3.530e-01  4.176e-01  -0.845 0.398000
jobmanagement      -7.302e-02  2.407e-01  -0.303 0.761602
jobretired          6.315e-01  3.112e-01   2.029 0.042454 *
`jobself-employed` -1.812e-01  3.533e-01  -0.513 0.608167
jobservices        -1.457e-01  2.729e-01  -0.534 0.593542
jobstudent          3.784e-01  3.750e-01   1.009 0.312958
jobtechnician      -1.926e-01  2.301e-01  -0.837 0.402496
jobunemployed      -6.395e-01  4.214e-01  -1.518 0.129138
jobunknown          5.207e-01  5.853e-01   0.890 0.373669
maritalmarried     -4.696e-01  1.743e-01  -2.694 0.007058 **
maritalsingle      -3.051e-01  2.038e-01  -1.497 0.134354
educationsecondary  8.011e-02  2.022e-01   0.396 0.691924
educationtertiary   3.208e-01  2.337e-01   1.373 0.169897
educationunknown   -4.210e-01  3.572e-01  -1.179 0.238561
default             5.446e-01  4.315e-01   1.262 0.206824
`Bank$balance`     -3.911e-06  1.749e-05  -0.224 0.823014
housing            -2.600e-01  1.381e-01  -1.883 0.059676 .
loan               -6.296e-01  2.000e-01  -3.149 0.001640 **
contacttelephone   -7.020e-02  2.327e-01  -0.302 0.762900
contactunknown     -1.416e+00  2.277e-01  -6.219 4.99e-10 ***
`Bank$day`          1.641e-02  8.161e-03   2.011 0.044362 *
monthaug           -3.081e-01  2.494e-01  -1.235 0.216655
monthdec            1.144e-01  6.573e-01   0.174 0.861784
monthfeb            2.022e-01  2.937e-01   0.688 0.491290
monthjan           -1.123e+00  3.816e-01  -2.944 0.003245 **
monthjul           -7.515e-01  2.498e-01  -3.008 0.002630 **
```

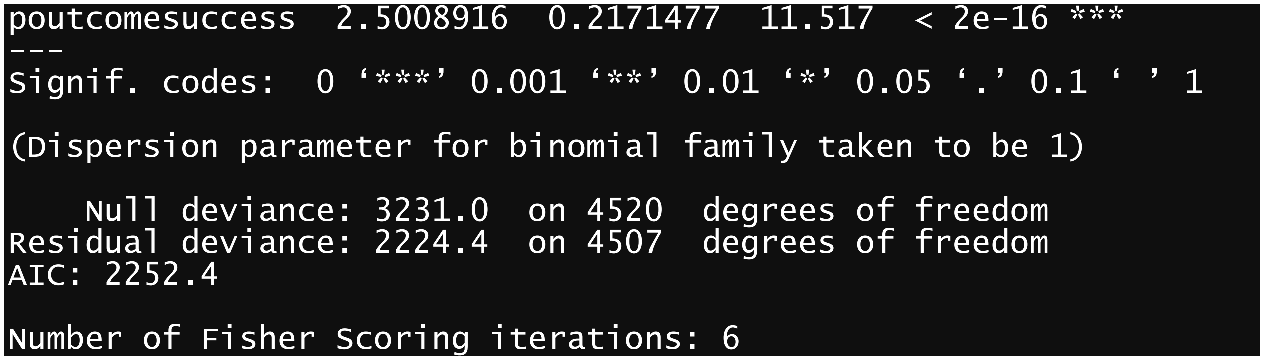


```
monthjun            5.542e-01  3.003e-01   1.845 0.065009 .  
monthmar            1.498e+00  3.901e-01   3.842 0.000122 ***
monthmay           -4.900e-01  2.340e-01  -2.094 0.036246 *  
monthnov           -8.430e-01  2.737e-01  -3.080 0.002072 ** 
monthoct            1.361e+00  3.300e-01   4.124 3.72e-05 ***
monthsep            6.572e-01  4.115e-01   1.597 0.110265    
duration            4.225e-03  2.020e-04  20.912  < 2e-16 ***
campaign           -7.042e-02  2.821e-02  -2.496 0.012549 *  
pdays              -9.791e-05  9.959e-04  -0.098 0.921684    
previous           -5.511e-03  3.818e-02  -0.144 0.885249    
poutcomeother       4.912e-01  2.692e-01   1.825 0.068019 .  
poutcomesuccess     2.445e+00  2.773e-01   8.818  < 2e-16 ***
poutcomeunknown    -1.216e-01  3.199e-01  -0.380 0.703822    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3231.0  on 4520  degrees of freedom
Residual deviance: 2173.7  on 4478  degrees of freedom
AIC: 2259.7

Number of Fisher Scoring iterations: 6
```

In order to simplify this, we employ the stepwise selection method with the BIC as the standard. And here is the result of selection.

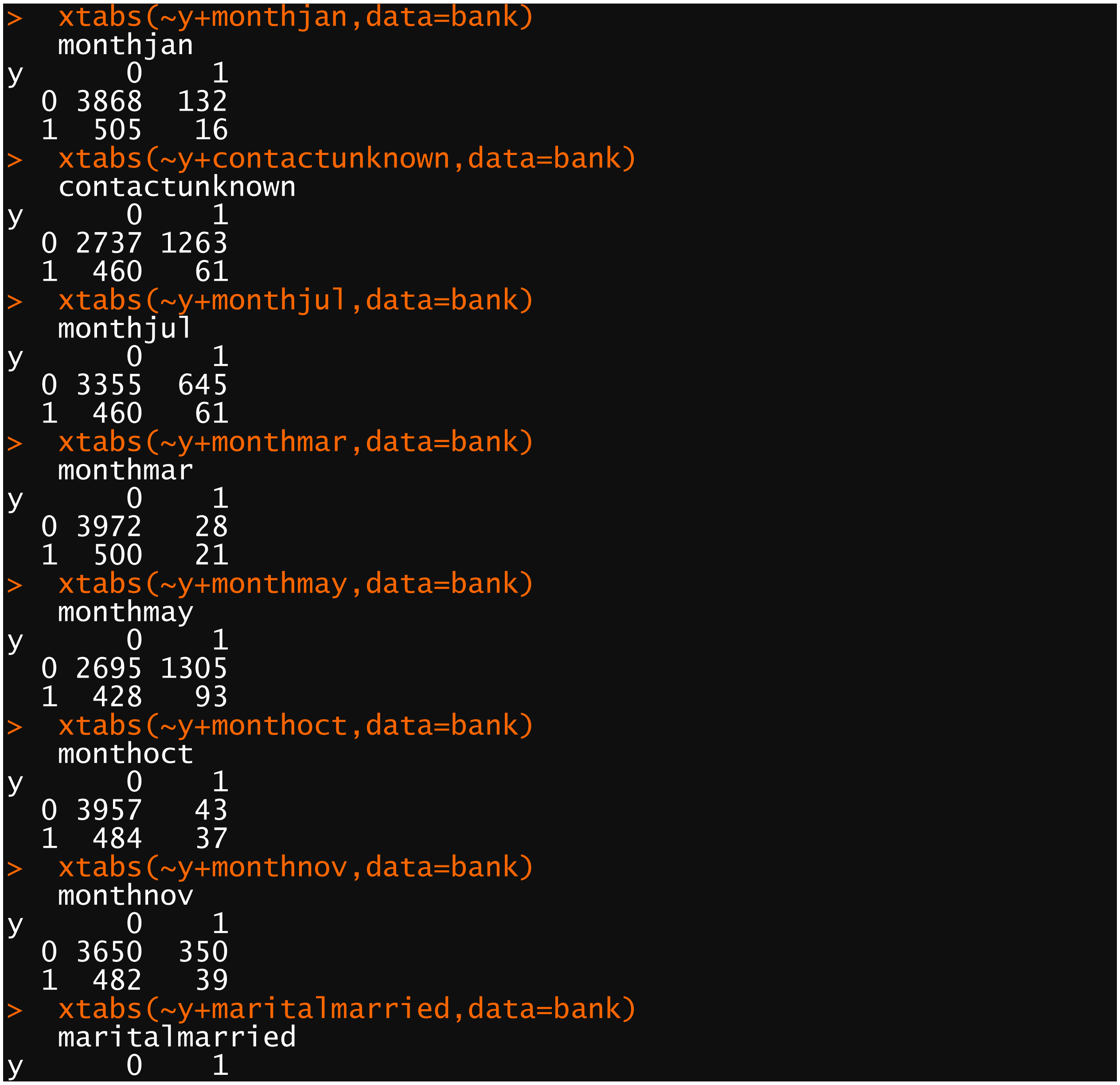


```
Call:
glm(formula = y ~ jobretired + maritalmarried + loan + contactun
known + 
    monthjan + monthjul + monthmar + monthmay + monthnov + month
oct + 
    duration + campaign + poutcomesuccess, family = "binomial", 
    data = bank)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-4.0803  -0.3972  -0.2651  -0.1672   3.0209  

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)    
(Intercept)     -2.6254372  0.1412887 -18.582  < 2e-16 ***
jobretired       0.7191126  0.2040986   3.523 0.000426 ***
maritalmarried  -0.4137360  0.1143798  -3.617 0.000298 ***
loan            -0.6306864  0.1953026  -3.229 0.001241 ** 
contactunknown  -1.2605895  0.1806082  -6.980 2.96e-12 ***
monthjan        -0.9397973  0.3360726  -2.796 0.005167 ** 
monthjul        -0.8058806  0.1813340  -4.444 8.82e-06 ***
monthmar         1.5364708  0.3438877   4.468 7.90e-06 ***
monthmay        -0.7157138  0.1565216  -4.573 4.82e-06 ***
monthnov        -0.8209415  0.2181389  -3.763 0.000168 ***
monthoct         1.4407692  0.2791453   5.161 2.45e-07 ***
duration         0.0041161  0.0001967  20.929  < 2e-16 ***
campaign        -0.0741598  0.0268909  -2.758 0.005819 ** 
```

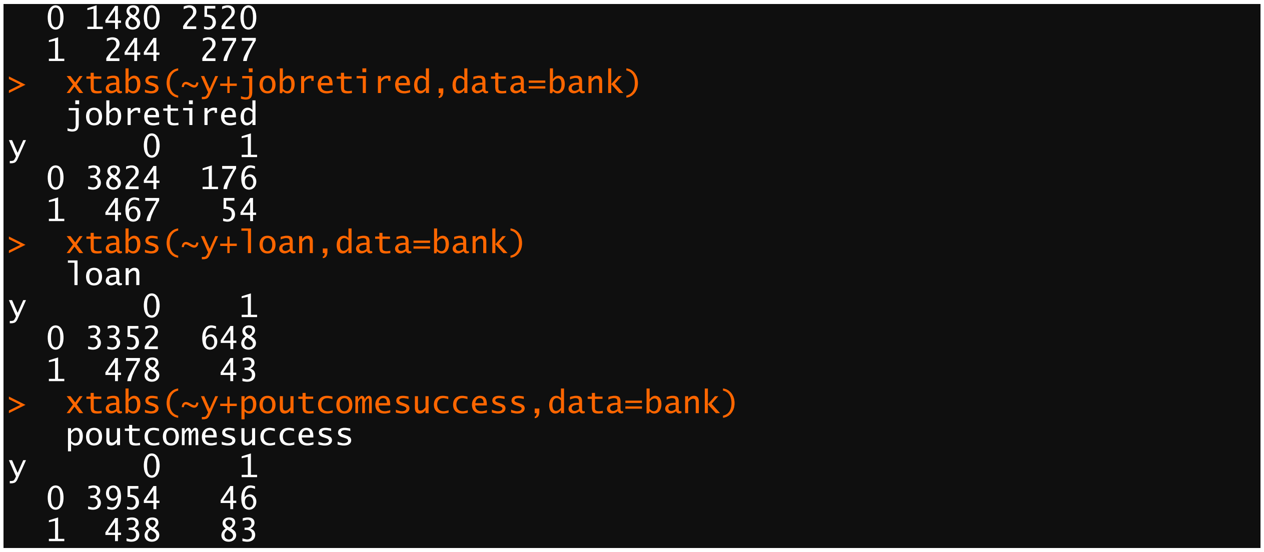


```
poutcomesuccess  2.5008916  0.2171477  11.517  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3231.0  on 4520  degrees of freedom
Residual deviance: 2224.4  on 4507  degrees of freedom
AIC: 2252.4

Number of Fisher Scoring iterations: 6
```

Then we are going to check if there is any complete separation or quasi-complete separation here. After we check all these dummy variables, we do not find any complete separation or quasi-complete separation. Here is the proof.

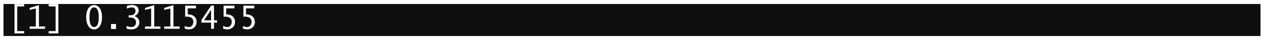


```
>  xtabs(~y+monthjan,data=bank)
   monthjan
y      0    1
  0 3868  132
  1  505   16
>  xtabs(~y+contactunknown,data=bank)
   contactunknown
y      0    1
  0 2737 1263
  1  460   61
>  xtabs(~y+monthjul,data=bank)
   monthjul
y      0    1
  0 3355  645
  1  460   61
>  xtabs(~y+monthmar,data=bank)
   monthmar
y      0    1
  0 3972   28
  1  500   21
>  xtabs(~y+monthmay,data=bank)
   monthmay
y      0    1
  0 2695 1305
  1  428   93
>  xtabs(~y+monthoct,data=bank)
   monthoct
y      0    1
  0 3957   43
  1  484   37
>  xtabs(~y+monthnov,data=bank)
   monthnov
y      0    1
  0 3650  350
  1  482   39
>  xtabs(~y+maritalmarried,data=bank)
   maritalmarried
y      0    1
```



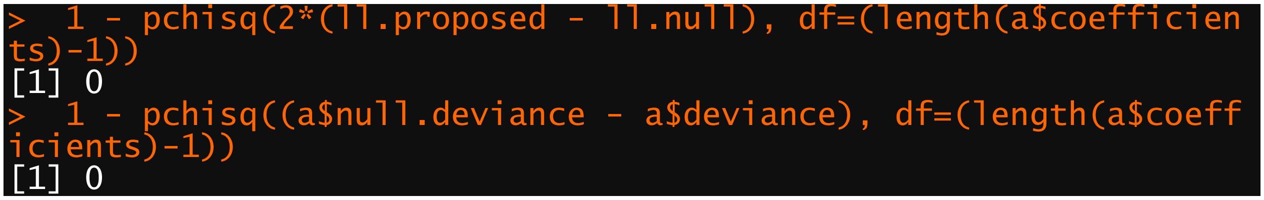
```
  0 1480 2520
  1  244  277
>  xtabs(~y+jobretired,data=bank)
   jobretired
y      0    1
  0 3824  176
  1  467   54
>  xtabs(~y+loan,data=bank)
   loan
y      0    1
  0 3352  648
  1  478   43
>  xtabs(~y+poutcomesuccess,data=bank)
   poutcomesuccess
y      0    1
  0 3954   46
  1  438   83
```

Then we compute the McFadden’s pseudo R^2 using attributes of the logistic regression. Here is the outcome.



```
[1] 0.3115455
```

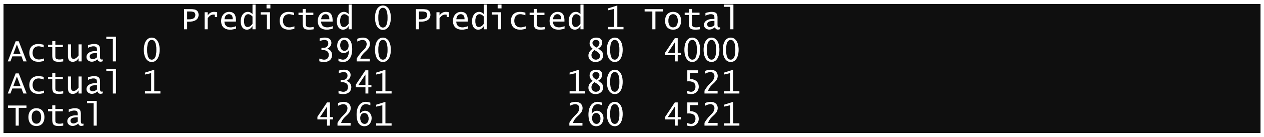
After that, we compute the p values of overall fitting test. Here is the result.



```
>  1 - pchisq(2*(ll.proposed - ll.null), df=(length(a$coefficien
ts)-1))
[1] 0
>  1 - pchisq((a$null.deviance - a$deviance), df=(length(a$coeff
icients)-1))
[1] 0
```

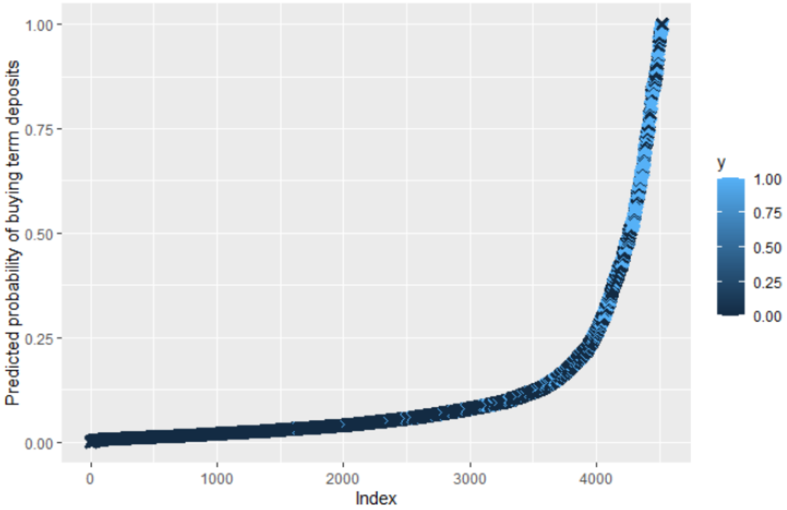
It appears that both p values indicate that the joint null hypothesis is rejected. That means the overall fitting is somewhat significant. And coefficients are significant.

Then we make predictions based on the resulting model selected by the BIC. And we also got the confusion matrix here. Here is the result.

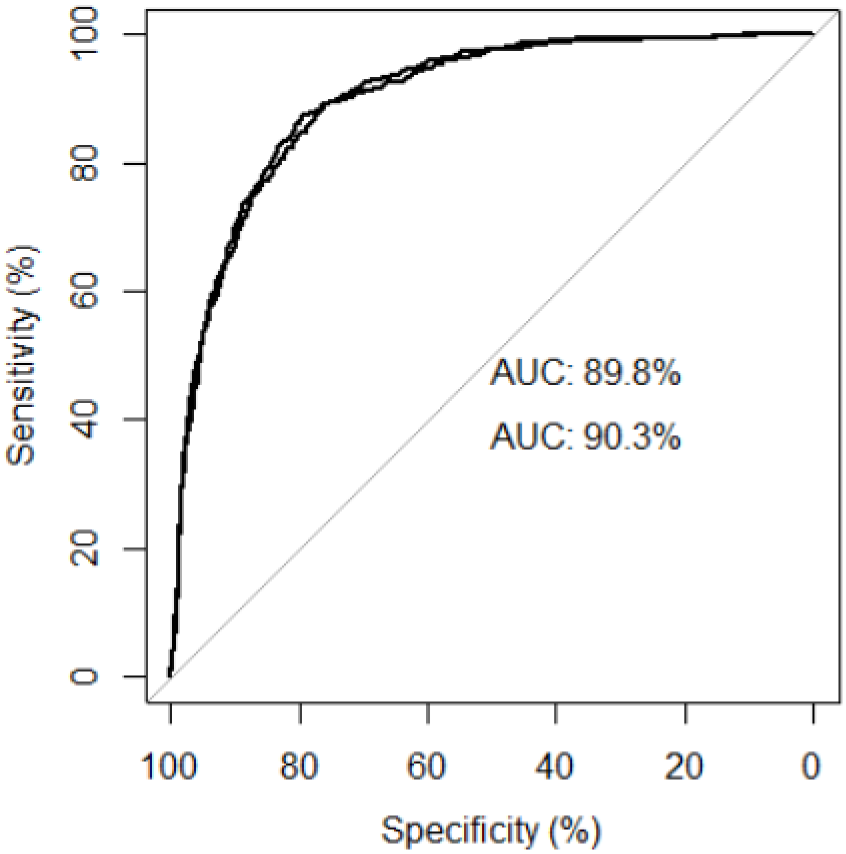


```
         Predicted 0 Predicted 1 Total
Actual 0        3920          80  4000
Actual 1         341         180   521
Total           4261         260  4521
```

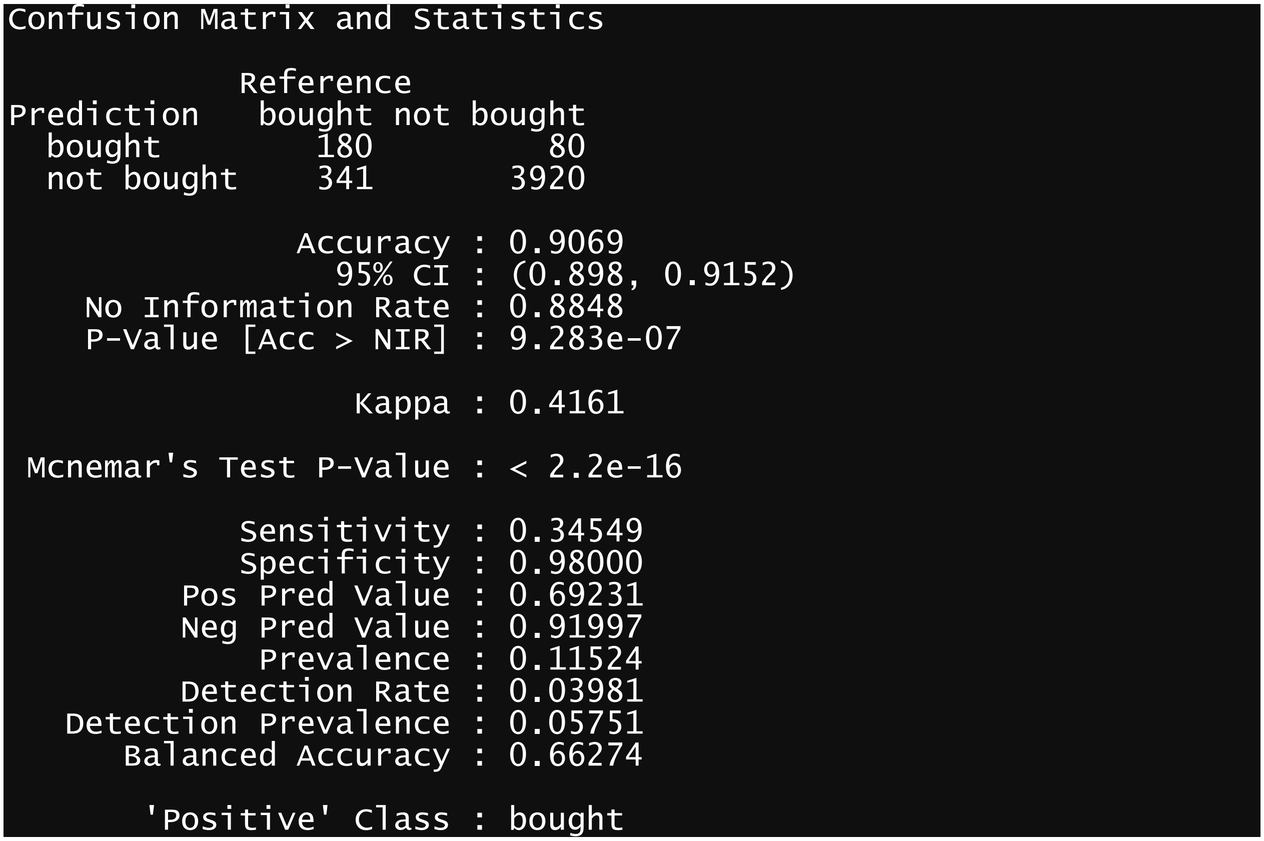
This matrix can describe the power of prediction to some extent. Then we tried to use graphs to show the prediction of probabilities about buying the term deposit product.



This plot shows the result of prediction and the truth. You can see that blue points mean that this client bought the product in truth while black dots means not buying in reality. While the x-axis represents the sample size, the y-axis means the probabilities of buying predicted from the model. As you can see, the model has some power in prediction. Later, we drew curves of precision and recall for both models above. Here is the outcome.



It turns out that both curves overlapped to some extent. That means their differences are minor while the former one has higher AUC OF 90.3%. But it also has higher BIC. On the other hand, the curve also indicates good performances of these models. Then we used another way to get the confusion matrix which is the same as that above. But this one has analysis with more details. And we set that probabilities over 0.5 indicate buying while probabilities below 0.5 mean not buying. Here it is.



```
Confusion Matrix and Statistics

            Reference
Prediction   bought not bought
  bought        180         80
  not bought    341       3920

               Accuracy : 0.9069
                 95% CI : (0.898, 0.9152)
    No Information Rate : 0.8848
    P-Value [Acc > NIR] : 9.283e-07

                  Kappa : 0.4161

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.34549
            Specificity : 0.98000
         Pos Pred Value : 0.69231
         Neg Pred Value : 0.91997
             Prevalence : 0.11524
         Detection Rate : 0.03981
   Detection Prevalence : 0.05751
      Balanced Accuracy : 0.66274

       'Positive' Class : bought
```